Learning and Generalization of Abstract Semantic Relations: Preliminary Investigation of Bayesian Approaches

Dawn Chen (sdchen@ucla.edu)
Department of Psychology

Hongjing Lu (hongjing@ucla.edu)
Departments of Psychology and Statistics

Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

A deep problem in cognitive science is to explain the acquisition of abstract semantic relations, such as antonymy and synonymy. Are such relations necessarily part of an innate representational endowment provided to humans? Or, is it possible for a learning system to acquire abstract relations from non-relational inputs of realistic complexity (avoiding hand-coding)? We present a series of computational experiments using Bayesian methods in an effort to learn and generalize abstract semantic relations, using as inputs pairs of specific concepts represented by feature vectors created by Latent Semantic Analysis.

Keywords: Bayesian inference; induction; generalization; abstract relations; machine learning; LSA

Introduction

An intelligent human adult can recognize that the concepts *day* and *night* are related in much the same way as *hot* and *cold*, but not in the same way as *day* and *hour*. This ability to appreciate abstract semantic relations is fundamental to analogical reasoning, and is arguably a core component of what is special about the human mind (Penn, Holyoak & Povinelli, 2007). But how are such abstract relations acquired? If they are learned, how this could be achieved is far from obvious. On the face of it, no perceptual or other features seem to be available to represent such abstract relations as antonymy, synonymy, or superordination. Almost by default, it might be assumed that abstract relations must be innate (Fodor, 1975).

Research on cognitive development has clearly established the phenomenon of a *relational shift* (Gentner & Rattermann, 1991), such that children process relations more effectively with increasing age. In particular, children move from a focus on global similarities of objects to similarities defined by specific dimensions, such as size or color (Smith, 1989; Smith & Sera, 1992). Less is known about the development of abstract relations that seem yet further divorced from perceptual similarity (see Halford, 1993). Analyses of corpora of child speech have identified systematic use of antonyms by children aged 2-5 years (Jones & Murphy, 2005). Children aged 6-7 years are more accurate in detecting the falsity of sentences such as *Some valleys are mountains* as compared to *Some valleys are lakes*, where the former sentence type contains an antonymous pair (Glass, Holyoak & Kossan, 1977), suggesting that some sense of antonymy is available prior to any formal instruction about this concept.

The Problem of Relation Learning

Regardless of whether abstract relations are learned or mature over the course of development, there is no doubt that adults can distinguish among instances of relations such as antonymy versus synonymy. In the present paper we pose the following computational problem: Given as inputs a modest number of pairs of concepts that instantiate an abstract relation (e.g., *day-night* and *hot-cold*, which instantiate antonymy), is it possible to extract a representation of the abstract relation that may then be used to accurately classify novel instantiations (e.g., *valley-mountain*)?

Most recent connectionist models of relation learning (e.g., Rogers & McClelland, 2008) have focused on the acquisition of small numbers of specific input-output pairs (e.g., “canary” + “can” → “fly”), but have not demonstrated the capacity to generalize to novel inputs dissimilar to the training items. In contrast, achieving such generalization is the central aim of our project. Moreover, an important constraint we imposed is that inputs to the learning system could not be hand-coded, as has been commonplace in the literature on computational models of analogy and relation learning. For example, Doumas, Hummel, and Sandhofer (2008) showed how structured relations corresponding to relative adjectives such as *bigger-than* can be extracted by bottom-up mechanisms given inputs consisting of unstructured feature vectors of objects. However, the modelers ensured that “size” features were present among the relatively small feature set defining the inputs, setting the stage for selecting these size features to form a part of the to-be-learned relational predicate. While perceptual relations may indeed be derived from the perceptual features of objects, this assumption is unwarranted for more abstract relations, for which hand-coding of features is even more problematic. In addition, realistic semantic representations would seem to require very large numbers of features, raising all the difficulties associated with search in a large representational space. Learning models that are developed for small, hand-tailored inputs at best postpone the challenges of "scaling up".Another approach to learning relations is to combine statistical techniques with structured representations. For example, Kemp and Tenenbaum (2008) showed how Bayesian techniques can operate on relational structures to learn relational systems such as hierarchies and linear orderings. The relational structures are provided to the system by including a grammar that generates possible structures. Although this approach may be appropriate for relations that have a well-defined logical structure known to the modeler, it is not clear that it can readily be extended to the full range of "messy" semantic relations. In addition, since the postulated grammar of relations is not itself learned, rather strong nativist assumptions remain.

Learning Relations from Unstructured Inputs

In this project, we have taken the tack of attempting to model the learning of abstract relations through essentially data-driven statistical learning, using Bayesian algorithms applied to large, unstructured input representations that we the modelers did not create. The raw inputs are vector representations of words, derived by Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). Such vectors, the product of singular value decomposition applied to lexical co-occurrence data from a large corpus of text, have proved extremely useful in many applications, often serving as good measures of semantic similarity of concepts (Wolf & Goldman, 2003). However, LSA vectors do not provide any direct basis for identifying abstract relations between concepts (although some modest success has been achieved by exploiting LSA vectors for relation words, such as *opposite*; Mangalath, Quesada & Kintsch, 2004). Related machine-learning algorithms have had some success in solving relational analogies by working directly from co-occurrence data for word combinations found in a large corpus of text (Turney & Littman, 2005). However, our goal is different in that we aim to model learning of relational representations from the LSA vectors for a small (< 20) set of word pairs that instantiate each abstract relation. The task of learning relations from representations of simpler concepts bears at least some resemblance to the task a child might face in acquiring an abstract relation from a modest-sized set of examples that instantiate it.

For our present purpose, we do not assume that LSA provides anything like an optimal psychological representation of concepts (indeed, it has well-known and serious limitations, notably problems dealing with lexical ambiguity). However, by using LSA inputs we ensure that we have in no way tailored the inputs so as to "hand hold" the learning algorithms we test. Moreover, we do not assume that it is in fact possible to acquire human-like representations of abstract relations solely by data-driven learning. Rather, by pressing the limits of data-driven approaches, we may be able to identify more clearly what nativist assumptions may ultimately prove essential.

A General Framework for Relation Learning

Here we report a preliminary investigation of relation learning based on two variants of the same basic framework. Our goal is to learn an explicit representation of a relation from a training set, **S**, consisting of pairs of concepts that each instantiate the relation. We assume that a decision regarding whether a pair of concepts instantiates a particular relation *R* is determined by a representation that includes both the basic features of the input concepts and additional features that the model automatically derives from the basic features. The full input representation is comprised of the basic features of two concepts, **A** and **B**, which are represented by LSA vectors, and of derived features ***Φ***(**A**, **B**) computed from **A** and **B** (see Fig. 1). In this study the derived features included two types, product features **AB** = $\begin{bmatrix} A_1B_1 & A_2B_2 & \cdots & A_dB_d \end{bmatrix}$ and absolute difference features $|\mathbf{A}-\mathbf{B}| = \begin{bmatrix} |A_1-B_1| & |A_2-B_2| & \cdots & |A_d-B_d| \end{bmatrix}$, both defined across corresponding positions in the **A** and **B** vectors. The length of each type of derived vector is thus equal to the length of each basic vector, so that the total size of the input vector scales linearly with the number of basic features.

If we let **X** denote the full vector including basic and derived features, **X** = [**A**, **B**, ***Φ***(**A**, **B**)], then the computational goal of relation learning is to estimate the distribution of a corresponding weight vector **w** from a set of training pairs that share the same relation. That is, we calculate $P(\mathbf{w} \mid \mathbf{X}_S, \mathbf{R}_S = 1)$, where the subscript **S** indicates the set of training examples (the source) and $\mathbf{R}_S$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular pair of concepts instantiates the relation or not. The vector **w** constitutes the learned relational representation, which can be interpreted as attention weights reflecting the importance of the corresponding features in **X**. To test generalization of the learned relational representation, we test on new transfer pairs, denoted by the subscript *T*. The inference step needs to estimate the probability that a target pair shares the same relation as the training pairs, $P(R_T = 1 \mid \mathbf{X}_T, \mathbf{X}_S, \mathbf{R}_S = 1)$.

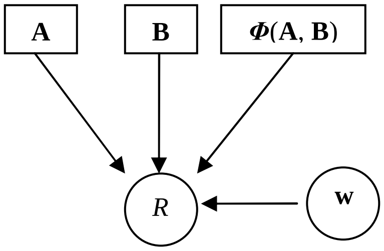

Figure 1: Graphical representation of the general framework. **A** and **B** denote two vectors of concept features (LSA inputs); ***Φ***(**A**, **B**) denotes derived features based on the two concepts, i.e., product features **AB** and absolute difference features |**A** – **B**|. Vector **w** represents the unknown relational weights that define *R*, and is learned using the training set of examples instantiating *R*.

The models we consider are both based on Bayesian logistic regression, as described by Silva, Airoldi and Heller (2007) and Silva, Heller and Gharamani (2007). Given a small set of word-pairs **S** that all instantiate a given abstract relation *R*, both models compute the posterior probability that $(\mathbf{A}_T, \mathbf{B}_T)$ is an example of the same relation,

$$P(R_T = 1 | \mathbf{X}_T, \mathbf{X}_S, \mathbf{R}_S = 1) = \int_{\mathbf{w}} P(R_T = 1 | \mathbf{X}_T, \mathbf{w}) P(\mathbf{w} | \mathbf{X}_S, \mathbf{R}_S = 1) \qquad (1)$$

where the likelihood is assessed using a logistic regression function to predict the probability of a word-pair instantiating a given relation,

$$P(R = 1 | \mathbf{X}, \mathbf{w}) = \text{logistic}(\mathbf{w}^{\mathrm{T}}\mathbf{X}) \qquad (2)$$

where $\text{logistic}(x) = \left(1 + e^{-x}\right)^{-1}$.

For the first model we consider (based directly on Silva et al., 2007), the posterior distribution for **w** is found by applying Bayes' rule using the prior distribution for **w** and the training word-pairs:

$$P(\mathbf{w} | \mathbf{X}_S, \mathbf{R}_S = 1) = \frac{P(\mathbf{R}_S = 1 | \mathbf{w}, \mathbf{X}_S) P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R}_S = 1 | \mathbf{w}, \mathbf{X}_S) P(\mathbf{w})} \qquad (3)$$

Because of the high dimensionality of the learning problem we are tackling, the choice of a good prior $P(\mathbf{w})$ is essential to the performance of any model. We investigated two kinds of priors, a simple empirical prior proposed by Silva and colleagues, and our own hierarchical model.

The Empirical Prior

Intuitively, our simple empirical prior distinguishes word-pairs that instantiate *any* of the to-be-learned relations from unrelated word-pairs. The empirical prior takes the form $P(\mathbf{w}) = N(\mathbf{w};\ \hat{\mathbf{w}}, \hat{\boldsymbol{\Sigma}})$, in which the sample mean estimate $\hat{\mathbf{w}}$ is by found by fitting a logistic regression classifier using maximum-likelihood estimation on a relatively small set of related word pairs (positive examples), and a larger set of unrelated word pairs (negative examples), reflecting the fact that most pairs of actual concepts do not instantiate any abstract relation. The covariance matrix $\hat{\boldsymbol{\Sigma}}$ for this empirical prior is calculated by

$$\left(\hat{\boldsymbol{\Sigma}}^{-1}\right) = c \cdot \left(\mathbf{X}^T \mathbf{M} \mathbf{X}\right) / N \qquad (4)$$

where *c* is a user-defined smoothing parameter set to twice the number of related pairs in the training samples, *N* is the total number of word pairs in the training set, and **X** is a matrix containing the features of all (related and unrelated) word pairs in the training set. **M** is a diagonal matrix with each entry defined as

$$\left(\mathbf{M}\right)_{ii} = \hat{p}(i)\left(1 - \hat{p}(i)\right) \qquad (5)$$

where $\hat{p}(i)$ is the MLE predicted probability of the *i*th word pair being related, given by Eq. (2).

The Hierarchical Prior

The above model computes its prior based on the observed data. This empirical prior uses all related pairs as members of the set of positive training cases, and numerous unrelated pairs as negative cases. An alternative empirical prior could be computed by considering pairs of a specific relation as positive examples and pairs instantiating other relations as negative examples. Although empirical priors are a sensible choice to facilitate inference in the high-dimensional space, the question of how the best data set for learning an empirical prior could be constructed remains unresolved.

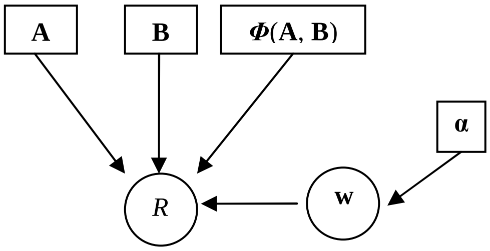

Figure 2: Graphical representation of hierarchical model. Distribution of **α** is determined by the hyperparameters that model the variance of the relational weight vector **w**. The other notations are the same as in Figure 1.

Here we explored a different approach, specifying a hierarchical prior on the distribution of the weight vector **w** (see Fig. 2). Specifically, the posterior distribution of **w** learned from training data is derived (replacing Eq. 3) by

$$P(\mathbf{w} | \mathbf{X}_S, \mathbf{R}_S = 1) = \int_{\boldsymbol{\alpha}} P(\mathbf{w} | \boldsymbol{\alpha}, \mathbf{X}_S, \mathbf{R}_S = 1) P(\boldsymbol{\alpha}) \qquad (6)$$

where vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots]$ determines the precision (the inverse variance) of each element of the weight vector $\mathbf{w}$. We use a conjugate prior distribution in the form of a Gamma distribution for α_i with two hyperparameters a_0 and b_0:

$$P(\alpha_i) \sim Gamma(\alpha;\ a_0, b_0) \qquad (7)$$

The individual prior for each element in vector $\mathbf{w}$ is assigned in the form of a normal distribution:

$$P(w_i | \alpha_i) \sim N(w_i;\ 0,\ \alpha_i) \qquad (8)$$

This normal distribution imposes a general prior that the value of w_i is centered at 0 (i.e., the *i*th feature dimension is not expected to be relevant in predicting whether a certain relation exists between the two words). However, the value of α_i controls the certainty about this prior belief. A low precision value makes the prior belief uninformative, whereas a high precision value imposes a strong bias that w_i is most likely 0. Accordingly, the hyperparameters play an important role in determining the relevance of feature dimensions in predicting the existence of a relation.

The other term in Eq. (6) can be derived by applying Bayes rule directly,

$$P(\mathbf{w} | \boldsymbol{\alpha}, \mathbf{X}_S, \mathbf{R}_S = 1) = \frac{P(\mathbf{R}_S = 1 | \mathbf{w}, \mathbf{X}_S) P(\mathbf{w} | \boldsymbol{\alpha})}{\int_{\mathbf{w}} P(\mathbf{R}_S = 1 | \mathbf{w}, \mathbf{X}_S) P(\mathbf{w} | \boldsymbol{\alpha})} \qquad (9)$$

The Inference Algorithm

Although the general framework of the relation learning models is straightforward, the inference step is non-trivial because the calculation of the normalization terms in Eqs. (3) and (9) and integrals in Eq. (6) are intractable, lacking analytic solutions. A sampling approach is impractical for dealing with high feature dimensionality. We therefore employed variational methods developed by Jaakkola and Jordan (2000) to obtain a closed-form approximation to the posterior distribution. Specifically, the variational method updates the mean of vector **w** and its covariance matrix **V** iteratively:

$$\begin{aligned}
&\mathbf{V}^{-1} = \mathbf{a}/\mathbf{b} + 2\sum \lambda(\boldsymbol{\xi}_\mathbf{n})\mathbf{x}_\mathbf{n}\mathbf{x}_\mathbf{n}^T,\\
&\mathbf{w} = \mathbf{V}\sum_n \mathbf{x}_\mathbf{n} / 2,\\
&\mathbf{a} = a_0 + 1/2, \qquad (10)\\
&\mathbf{b} = b_0 + \mathrm{E}_w(\mathbf{w}\mathbf{w}^T)/2,\\
&\boldsymbol{\xi}^\mathbf{2} = \mathbf{x}^T(\mathbf{V} + \mathbf{w}\mathbf{w}^T)\mathbf{x}.
\end{aligned}$$

Computational Experiments

The Training Set and Generalization Test

Table 1 shows some examples of pairs of concepts that we used to train and test the two models. We used four different relations: function, synonyms, linear ordering, and antonyms. For each relation, we chose 15-20 pairs that were examples of that relation to use as the training set. We will refer to pairs used for training as AB pairs. All pairs were selected from experimental materials used previously to form four-term verbal analogy problems, and for which LSA vectors (derived using the tasaALL corpus) were available. We selected pairs for which the cosine similarity between the words (based on their LSA vectors) was at least 0.1, aiming to exclude pairs that included highly ambiguous words (e.g., *gift-present* as an example of synonyms).

After learning representations of the abstract relations based on the AB pairs, the model was tested on a two-alternative forced-choice generalization task. For each test item, the model was asked to choose which of two alternative pairs instantiated a specified relation. We will refer to correct and incorrect options as CD and CD', respectively. For example, one item required the models to decide which pair instantiated antonymy, *shallow-deep* (CD) or *shallow-depth* (CD'). As this example suggests, the discrimination was quite subtle, as the C term was common to both options and the CD' pair also instantiated an abstract relation (but not the relation being queried). The words used in this generalization test did not overlap at all with the AB pairs used in training, but were selected according to the same general criteria. For each test problem, the models calculated the probability of CD and of CD' being examples of the relation, respectively, according to Eq. (1), and chose the pair with the higher probability as the answer. The percentage of test questions that each model answered correctly for each relation was calculated.

Table 1: Examples of word pairs used in the training sets and generalization tests (correct option on left).

Training pairs	Testing pairs
Function	
door-open	rabbit-hop vs. rabbit-bunny
sun-warm	cup-drink vs. cup-mug
zoo-animals	smile-happy vs. smile-frown
Synonyms	
liberty-freedom	car-auto vs. car-bus
huge-enormous	weak-feeble vs. weak-strong
forest-woods	sad-unhappy vs. sad-sadder
Linear ordering	
worse-worst	inch-foot vs. inch-length
kitten-cat	rain-downpour vs. rain-fall
tap-strike	pebble-rock vs. rock-mineral
Antonyms	
weak-strong	shallow-deep vs. shallow-depth
start-finish	float-sink vs. float-boat
slowly-quickly	find-lose vs. find-search

Simulation Details

Inputs for each word were LSA vectors of length 300. The LSA algorithm orders its features from highest to lowest in terms of their predictive power. Preliminary tests indicated that most of the information useful for our learning models was encoded in the first ten features of the LSA vectors. Accordingly, we used just these first ten features for each word as inputs. The full vector for a word pair included the basic and derived features, **X = [A, B, AB, |A – B|]**, with a total length of 40 features.

In the implementation of the model by Silva et al. (2007), the dataset for computing the empirical prior included all AB word pairs plus a large number (>3500) of unrelated word pairs. Each unrelated word pair was weighted by approximately the ratio of the total population of unrelated word pairs to the number of unrelated word pairs that were sampled. After obtaining the prior, the model employed variational methods to compute the posterior distribution for **w** using the AB training pairs for each relation separately.

In the simulation of our hierarchical model, the values of hyperparameters (a_0, b_0) were searched separately for each relation to maximize generalization performance.

To provide baselines for evaluating the two Bayesian learning models, we applied three simpler methods of judging the correct relational alternative. First, we calculated the mean cosine distances of the correct alternative and its foil to the training set using “raw” LSA vectors, i.e., using only the basic features [**A**, **B**] over all 300 dimensions of the LSA vector for each word in a pair (yielding 600 features total). Specifically, we computed the average of cosine distances between a CD pair and all AB pairs in the training set, and for the corresponding CD' pair and all AB pairs. The baseline decision for the discrimination task was determined by which pair yielded the closer cosine distance. The performance of this method informs us about the amount of information that "raw" LSA vectors provide for the four abstract relations of interest.

Second, we used an additional cosine distance measure defined over the same feature vectors as those used by the Bayesian models, i.e., the **X** vectors, which included the first ten features of the LSA vector for each word, plus the corresponding derived features.

Third, we examined the performance of simple logistic regression (which obtains the relational representation **w** through maximum-likelihood estimation) using the first ten LSA dimensions and the full set of derived features.

Results and Discussion

The five modeling methods were evaluated on nine different sets of training pairs and testing pairs. Each set was randomly chosen from the analogy problems available to us. Mean proportion correct over the nine different training/test sets for each of the methods described above is shown in Fig. 3. Overall, the Bayesian model incorporating the hierarchical prior yielded the best generalization performance for all four relations, and in each case was reliably more successful than any of the three baseline models. The proportions correct for the hierarchical model were .78 for function, .72 for synonyms, .86 for linear ordering, and .66 for antonyms. In general, the generalization performance for the Bayesian models was best for linear ordering and weakest for antonymy. It should be noted that the linear ordering relation can be viewed as a generalization of the type of specific comparative relation (e.g., "larger than") to which the learning model proposed by Doumas et al. (2007) has been applied.

The Importance of the Prior

The improvement in generalization performance of the Bayesian models over the MLE logistic regression model illustrates the importance of the prior distribution on the relational weights **w**. This result suggests the possibility that children may also benefit from prior knowledge, either innate or acquired through previous experience, when learning new abstract relations. They may, for example, first learn to distinguish related or generally similar concepts from unrelated concepts before discriminating among more specific relations. Future experiments could explore the kinds of prior training that best aid human learning of new abstract relations, and compare the results with model performance using different priors.

The superior generalization of the Bayesian model using the hierarchical prior compared with the model using the empirical prior indicates that learning can be further improved by introducing a more effective prior. Using the general prior knowledge obtained by contrasting related and unrelated relations is a sensible choice in the applications on which Silver et. al. (2007) focused. However, this empirical prior may not be sufficient to provide informative guidance for inferences in the high dimensional space created using LSA inputs. Adopting a hierarchical prior increases learning power by incorporating soft constraints on the relational representation, **w**, and its associated uncertainty**.**

Why are Antonyms so Hard?

The fact that the Bayesian models performed relatively poorly on antonyms warrants further analysis. It should be noted that for antonyms only, the cosine distance method based on 300 LSA dimensions (with basic features only) outperformed cosine distance based on 10 LSA dimensions and the full set of derived features. This finding raises the possibility that finding a good representation for antonymy

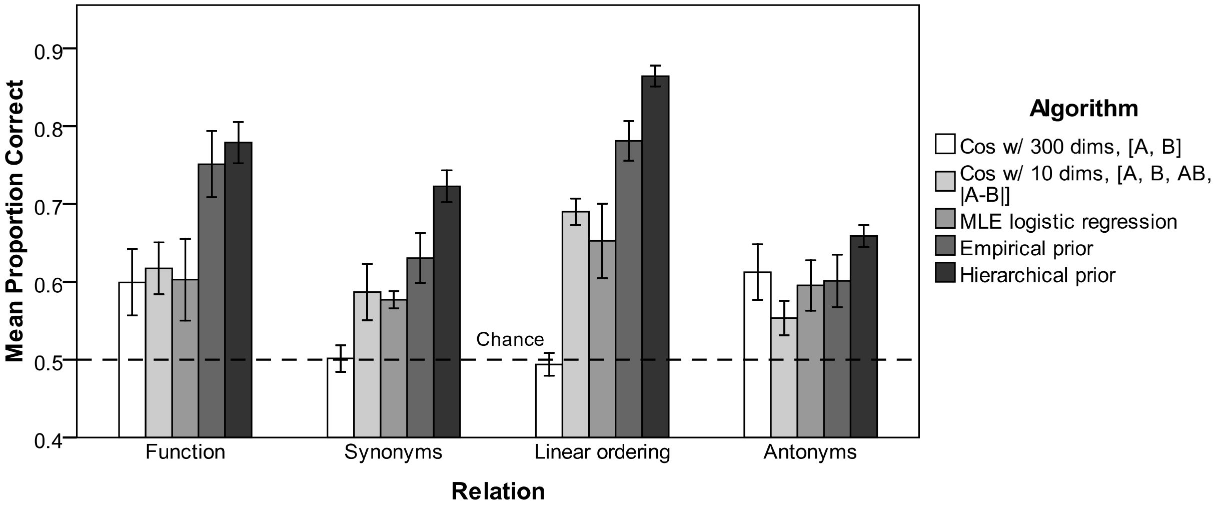

Figure 3: Simulation results. Prediction accuracy for generalization of relations in the two-alternative forced-choice relation-discrimination task. Error bars represent 1 standard error of the mean, based on 9 random samples of training/test items.

may require attention to more feature dimensions than is the case for the other relations. Another possible reason for their greater difficulty is that antonyms are usually very similar concepts that are dissimilar in only a few aspects (e.g., both *love* and *hate* can be used as a noun as well as a verb, and are strong emotions that one sentient being can have about another). Moreover, the aspects or dimensions on which antonymous concepts differ vary from one pair to another (e.g., *love-hate* vs. *black-white*). The shifting relevance of features makes learning a good representation for antonyms challenging, especially using a method that learns weight distributions over a fixed set of features.

Conclusions

We investigated the possibility that abstract semantic relations can be learned at least in part by purely data-driven statistical techniques applied to concept pairs represented by unstructured feature vectors. By using LSA vectors as inputs we avoided any hand-coding of semantics or relational structure, while assuring that inputs were of realistic complexity. Compared to baseline performance (inference based on cosine similarity of test options to the training set and MLE logistic regression), two models of relation learning based on Bayesian logistic regression achieved higher overall performance on a transfer test requiring discrimination between learned relations instantiated entirely by new concepts. The more successful of the two models incorporated hierarchical priors.

Neither model approached perfect performance on transfer problems. However, considering the small size of the training set (less than 20 examples of each relation), the total absence of overlap between training and test items, and the relatively subtle discrimination of relations required on the generalization test, these preliminary findings are encouraging. Further exploration of statistical approaches to learning abstract semantic relations appears to be warranted.

Acknowledgments

We thank Walter Kintsch, Tom Landauer and Praful Mangalath at the University of Colorado Institute of Cognitive Science for kindly providing us with LSA vectors. This research was funded by a University Fellowship and a Chancellor’s Prize from the Graduate Division at the University of California, Los Angeles, and by ONR grant N000140810186.

References

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development*. London: Cambridge University Press.

Glass, A. L., Holyoak, K. J., & Kossan, N. E. (1977). Children’s ability to detect semantic contradictions. *Child Development*, *48*, 279-283.

Halford, G. S. (1993). *Children’s understanding: The development of mental models.* Hillsdale, NJ: Erlbaum.

Jaakkola, T. S., & Jordan, M. I. (1999). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*, 25-37.

Jones, S., & Murphy, M. L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics*, *10*, 401-422.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA*, *105*, 10687-10692.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Mangalath, P., Quesada, J., & Kintsch, W. (2004). Analogy-making as predication using relational information and LSA vectors. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin’s mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109-178.

Rogers, T. T., & McClelland, J. L. (2008). Précis of *Semantic cognition: A parallel distributed processing approach*. *Behavioral and Brain Sciences*, *31*, 689-714.

Silva, R., Airoldi, A., & Heller, K. (2007). *Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning* (Tech. Rep. GCNU TR 2007-001). London: University College London, Gatsby Computational Neuroscience Unit.

Silva, R., Heller, K., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, UK: Cambridge University Press.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*, 99–142.

Turney, P., & Littman, M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*, 251–278.

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behaviour Research Methods*, *35*, 22-31.